Comparing Forecast Skill

Timothy DelSole

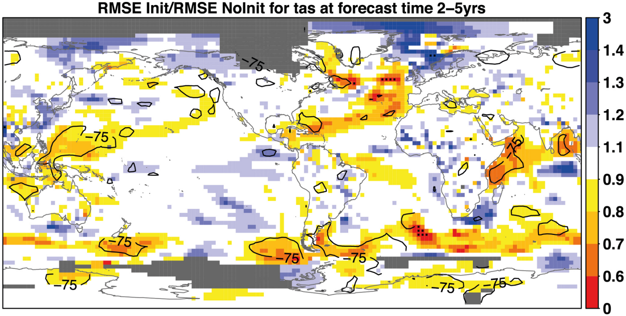
George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

October 6, 2015

collaborator: Michael Tippett

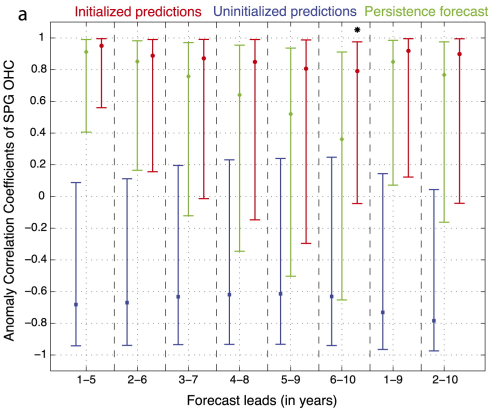
Is one forecast better than another?

- Operational centers want to know which prediction system to use.
- Modelers want to know if prediction system changes improved skill.
- Users want to know how prediction systems have performed in past.
- Scientists want to know if improvement is from: model resolution, initialization, model physics, ensemble format, recalibration.



Ratio of root mean square error of initialized over uninitialized decadal hindcasts. Dots indicate where the ratio is significantly above or below 1 with 90% confidence using a two-sided F-test.

IPCC AR5 WG1 fig. 11.4



Anomaly correlations of the North Atlantic Subpolar Gyre OHC anomalies (circle). The bar indicates the two-sided 90% confidence interval using Fishers z transform.

Msadek et al., 2014, J. Climate

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

$$F = \frac{s_1^2}{s_2^2}.$$

Theorem: If samples are independent and identically distributed as a Gaussian, then

$$F \sim F_{\nu_1,\nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

$$F = \frac{s_1^2}{s_2^2}.$$

Theorem: If samples are **independent** and identically distributed as a Gaussian, then

$$F \sim F_{\nu_1, \nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

Errors Computed over Same Period are not Independent

$$\begin{aligned} o &= s_o + n_o && \text{observation} \\ f_1 &= s_1 + n_1 && \text{forecast 1} \\ f_2 &= s_2 + n_2 && \text{forecast 2} \end{aligned}$$

The covariance between forecast errors is

$$\underbrace{\operatorname{cov}[f_1 - o, f_2 - o]}_{\text{cov. of errors}} = \underbrace{\operatorname{cov}[s_1 - s_o, s_2 - s_o]}_{\text{correlated signal errors}} + \underbrace{\operatorname{var}[n_o]}_{\text{obs. noise var.}}$$

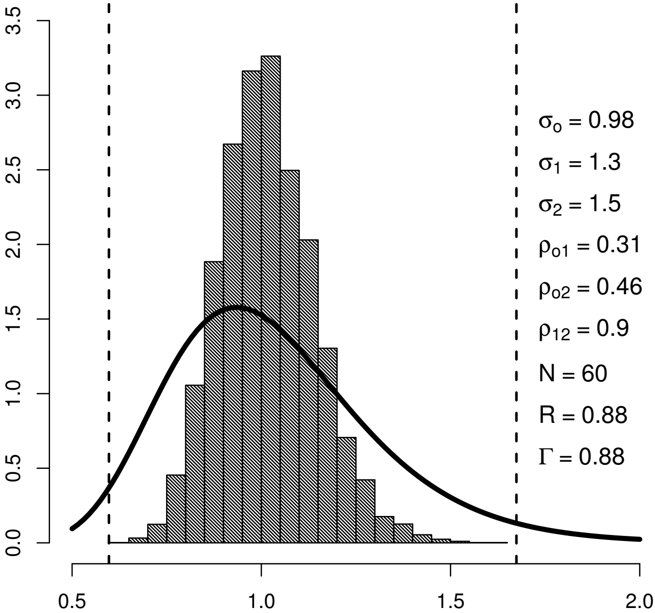
Since the covariance never vanishes, errors are not independent.

Correlated Errors Tend to be Closer to Each Other

$$\text{ratio of errors} = \frac{\sum_n (f_1 - o)^2}{\sum_n (f_2 - o)^2} = \frac{\sum_n \left(s_1 - s_o + n_1 - n_o\right)_n^2}{\sum_n \left(s_2 - s_o + n_2 - n_o\right)_n^2}$$

Top and bottom sums involve n_o, so the ratio will be closer to unity than would be the case if all terms were independent.

Ratio of Mean Square Errors



Response from IPCC

The reviewer is right ... However, no methodology appropriate for the decadal prediction problem is yet available. The methodology used is described in Doblas-Reyes et al (2013) ... intends to be more conservative than the typical formula described in the text book of von Storch and Zwiers.

Impact of Correlated Errors

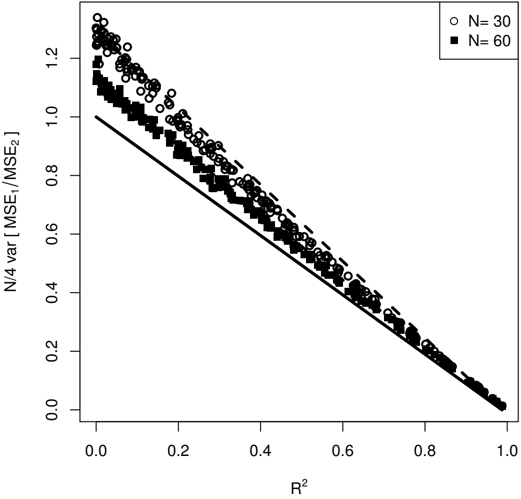
For large sample size N,

$$\text{var}\left[\frac{MSE_1}{MSE_2}\right] \approx \frac{4}{N}\left(1 - R^2\right),$$

where R is the correlation between forecast errors.

DelSole and Tippett, 2014: Comparing Forecast Skill. MWR.

Error ratio for randomly selected parameters of an idealized forecast/observation system



Test Equality of Correlations ($\rho_1 = \rho_2$)

Transform to a Gaussian variable using Fisher Z-transformation:

$$z \equiv \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) .$$

Apply standard t-test for a difference in means: i.e., test $z_1 = z_2$

Test Equality of Correlations ($\rho_1 = \rho_2$)

Transform to a Gaussian variable using Fisher Z-transformation:

$$z \equiv \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right).$$

Apply standard t-test for a difference in means: i.e., test $z_1 = z_2$

Assumes the two correlations computed from **independent** data.

Correlations Reduce the Variance of Correlation Differences

$$\mathrm{var}[z_1 - z_2] = \frac{2}{N-3}(1 - \Gamma) \qquad \text{where} \quad \Gamma = \mathrm{cor}[z_1, z_2]$$

DelSole and Tippett, 2014: Comparing Forecast Skill. MWR.

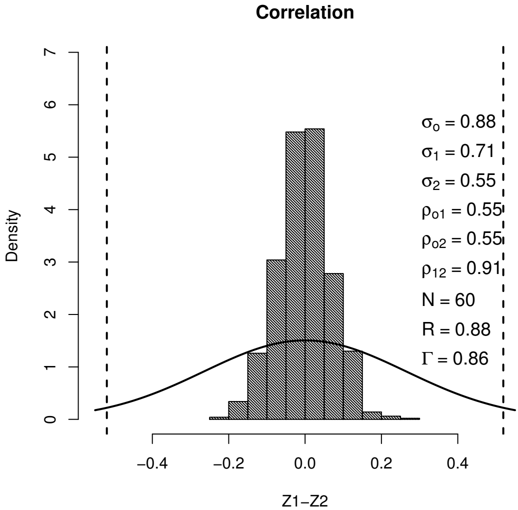
Correlations Reduce the Variance of Correlation Differences

$$\mathrm{var}[z_1 - z_2] = \frac{2}{N-3}(1-\Gamma) \qquad \text{where} \quad \Gamma = \mathrm{cor}[z_1, z_2]$$

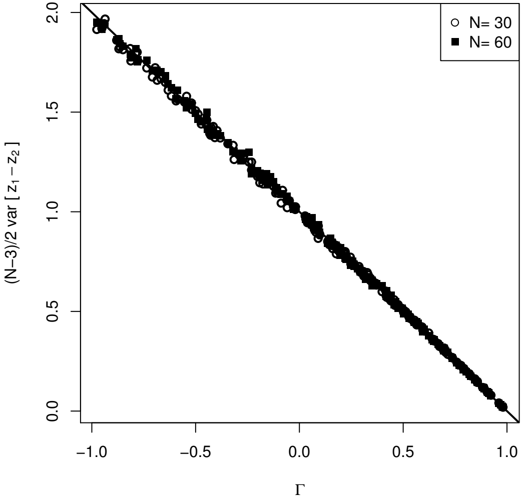
For large N and Gaussian distributions,

$$\Gamma = \frac{\rho_{12}\left(1-\rho_{o1}^2-\rho_{o2}^2\right) - \rho_{o1}\rho_{o2}\left(1-\rho_{o1}^2-\rho_{o2}^2-\rho_{12}^2\right)/2}{\left(1-\rho_{o1}^2\right)\left(1-\rho_{o2}^2\right)}.$$

DelSole and Tippett, 2014: Comparing Forecast Skill. MWR.


Correlation
Density
0
1
2
3
4
5
6
7
−0.4
−0.2
0.0
0.2
0.4
Z1−Z2
$\sigma_o = 0.88$
$\sigma_1 = 0.71$
$\sigma_2 = 0.55$
$\rho_{o1} = 0.55$
$\rho_{o2} = 0.55$
$\rho_{12} = 0.91$
N = 60
R = 0.88
$\Gamma = 0.86$

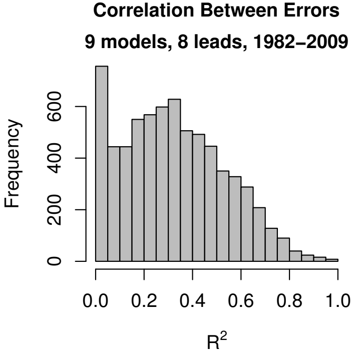
Variance of Difference in Zs for randomly selected parameters of an idealized forecast/observation system



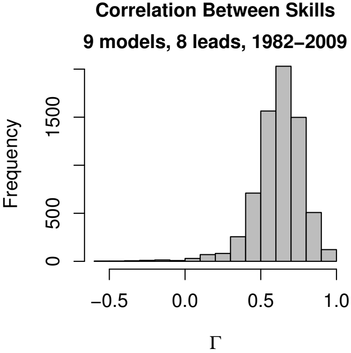
North American Multi-Model Ensemble

- Hindcasts initialized every month from 1982-2010
- At least 8 month lead
- Analyze NINO3.4
- Separate climatologies for 1982-1998 and 1999-2010
- Verification: OISST

model	ensemble size
CMC1-CanCM3	10
CMC2-CanCM4	10
COLA-RSMAS-CCSM3	6
GFDL-CM2p1	10
NASA-GMAO	10
NCEP-CFSv1	10
NCEP-CFSv2	10



Correlation Between Errors
9 models, 8 leads, 1982–2009



Correlation Between Skills
9 models, 8 leads, 1982–2009

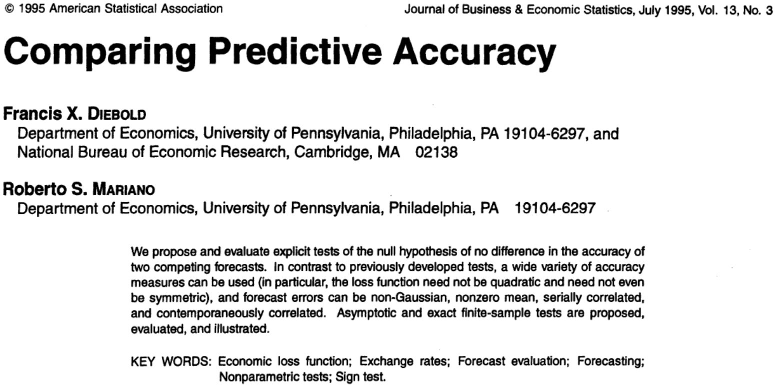
Skill estimates tend to be correlated in seasonal forecasting.

Summary

1. Commonly used tests for skill differences are not valid if skills are computed using a common set of observations.
2. These tests do not account for correlations between skill estimates.
3. These tests are biased toward indicating no difference in skill.
4. The bias can be characterized by a few parameters that can be estimated from data.
5. The bias is substantial for typical seasonal forecasts.

Familiar tests wrongly judge that differences in forecast skill are insignificant.

What **IS** the proper way to compare forecast skill?



© 1995 American Statistical Association

Journal of Business & Economic Statistics, July 1995, Vol. 13, No. 3

Comparing Predictive Accuracy

Francis X. Diebold
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, and
National Bureau of Economic Research, Cambridge, MA 02138

Roberto S. Mariano
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297

We propose and evaluate explicit tests of the null hypothesis of no difference in the accuracy of two competing forecasts. In contrast to previously developed tests, a wide variety of accuracy measures can be used (in particular, the loss function need not be quadratic and need not even be symmetric), and forecast errors can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Asymptotic and exact finite-sample tests are proposed, evaluated, and illustrated.

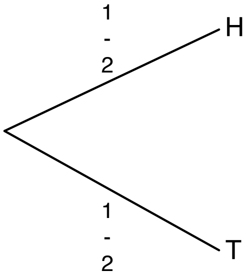
KEY WORDS: Economic loss function; Exchange rates; Forecast evaluation; Forecasting; Nonparametric tests; Sign test.

“The literature contains literally thousands of forecast-accuracy comparisons; almost without exception, point estimates of forecast accuracy are examined, with no attempt to assess their sampling uncertainty.”

Diebold and Mariano (1995)

Skill of Single Events

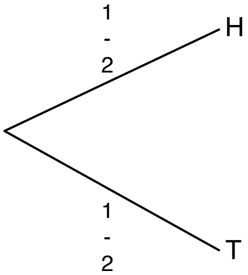
Identify Events When Forecast H has more skill than Forecast T.



Null hypothesis: probability that H has more skill than T is 50/50.

Skill of Single Events

Identify Events When Forecast H has more skill than Forecast T.

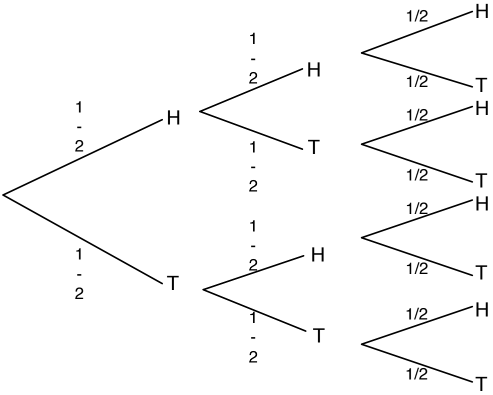


Null hypothesis: probability that H has more skill than T is 50/50.

- No caveats about independence.
- No assumptions about distribution of forecast errors.
- No restrictions on the criterion for deciding skill.

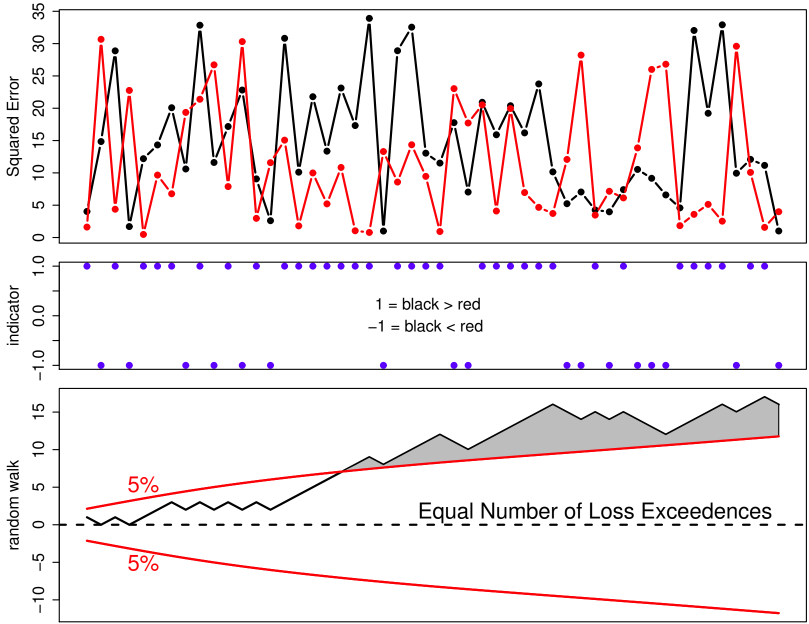
Random Walk Test

Identify Events When Forecast H has more skill than Forecast T.



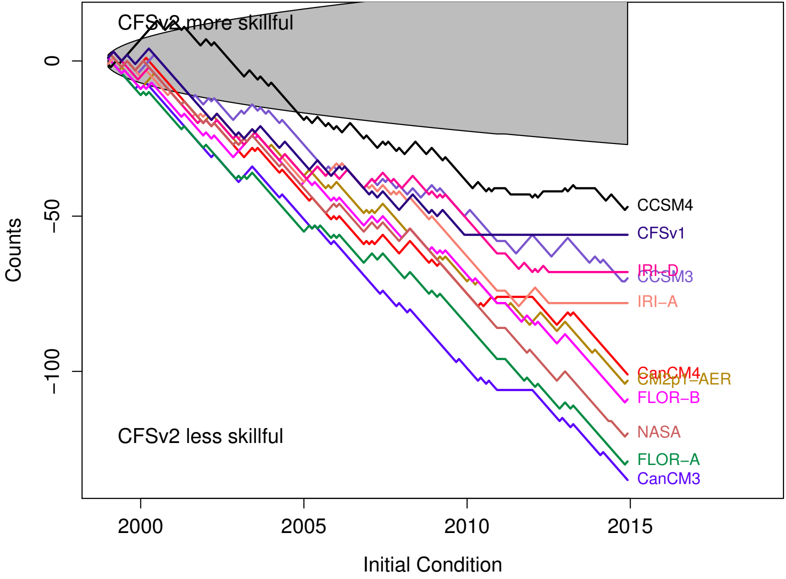
Null hypothesis: Counts follow a binomial distribution with p=1/2.

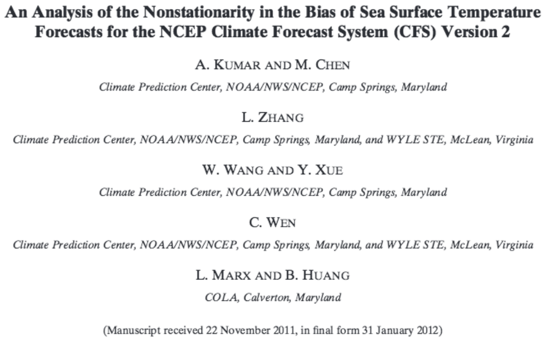
Visualizing the Verification: Random Walks



Remove climatologies based on 1982-1998 training

Monthly Mean NINO3.4 Forecasts by CFSv2
1982–1998 CLIM; lead= 2.5; alpha= 5%





An Analysis of the Nonstationarity in the Bias of Sea Surface Temperature Forecasts for the NCEP Climate Forecast System (CFS) Version 2

A. KUMAR AND M. CHEN

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland

L. ZHANG

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland, and WYLE STE, McLean, Virginia

W. WANG AND Y. XUE

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland

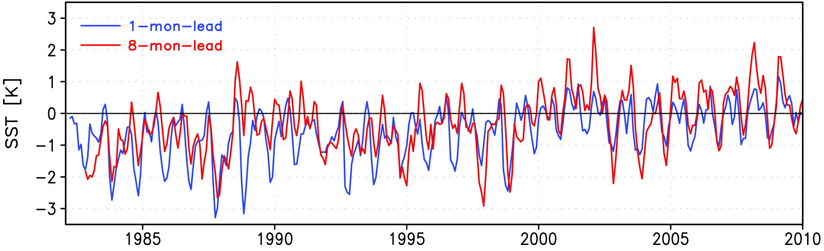
C. WEN

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland, and WYLE STE, McLean, Virginia

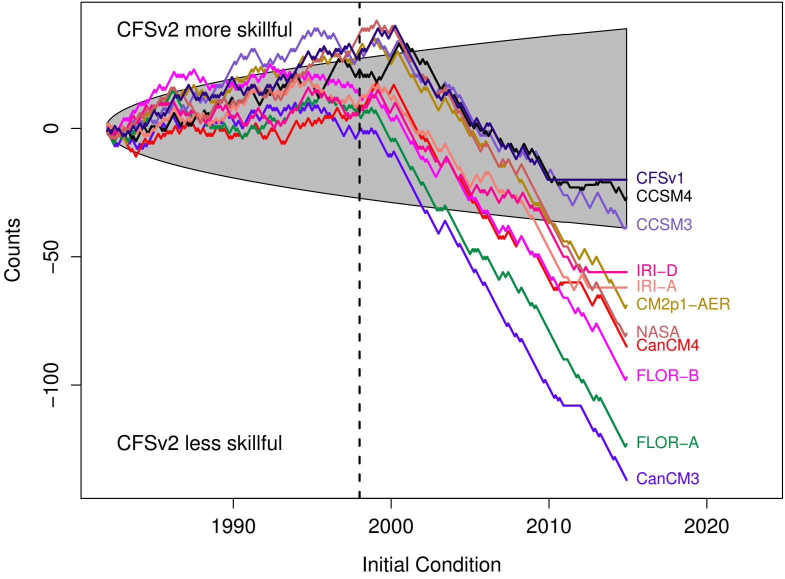
L. MARX AND B. HUANG

COLA, Calverton, Maryland

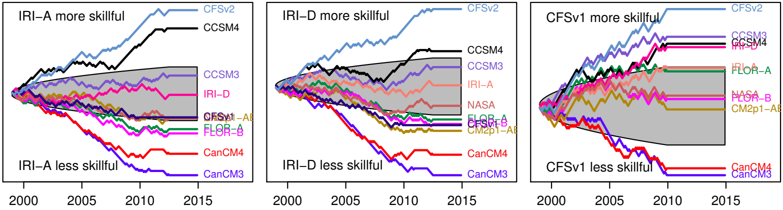
(Manuscript received 22 November 2011, in final form 31 January 2012)



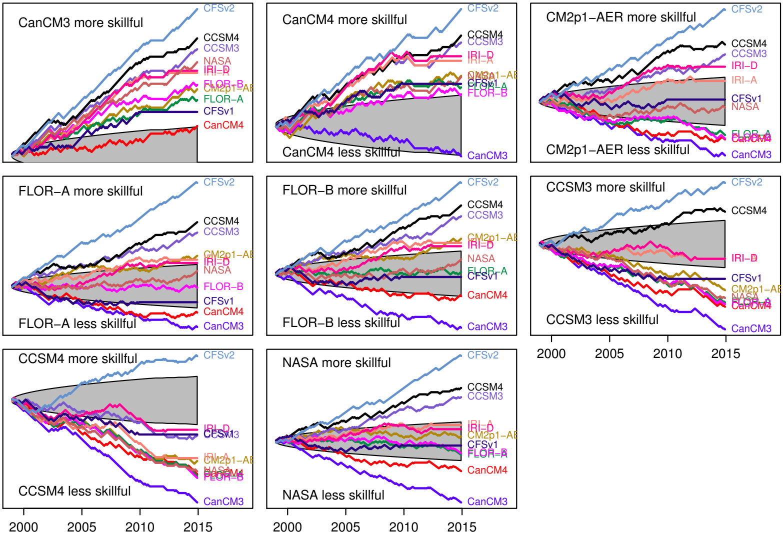
Monthly Mean NINO3.4 Forecasts by CFSv2
1982–1998 CLIM; lead= 2.5; alpha= 5%



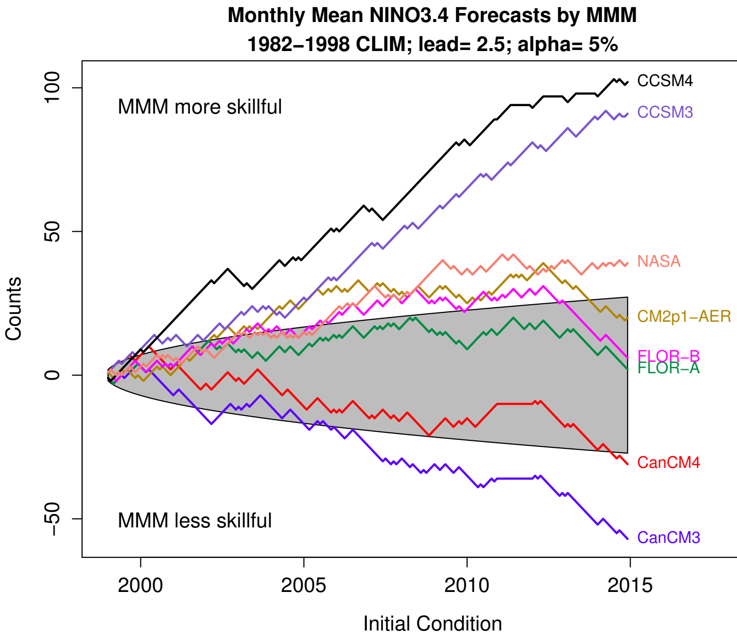
Comparison of Monthly Mean NINO3.4 Hindcasts of NMME Models
1982–1998 CLIM; lead= 2.5; alpha= 5%



Comparison of Monthly Mean NINO3.4 Hindcasts of NMME Models
1982–1998 CLIM; lead= 2.5; alpha= 5%



Multimodel Mean



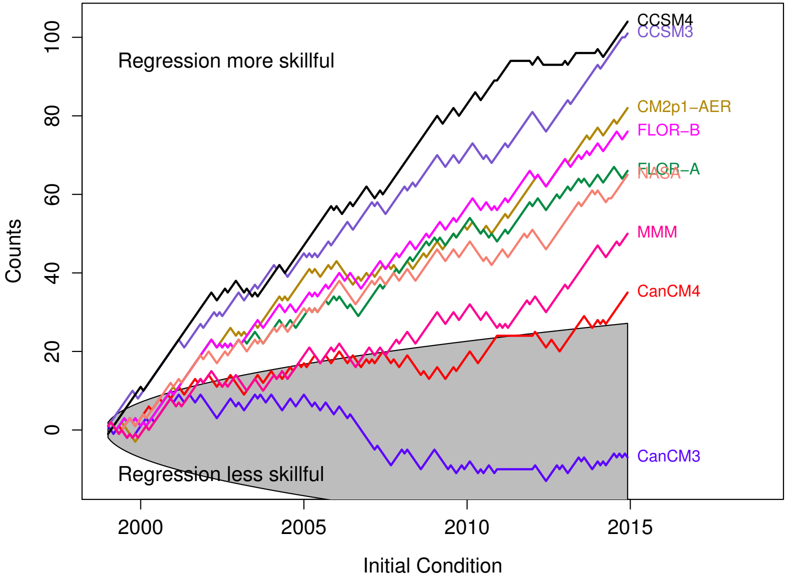
Monthly Mean NINO3.4 Forecasts by MMM
1982–1998 CLIM; lead= 2.5; alpha= 5%
MMM more skillful
MMM less skillful
CCSM4
CCSM3
NASA
CM2p1–AER
FLOR–B
FLOR–A
CanCM4
CanCM3
Counts
100
50
0
−50
2000
2005
2010
2015
Initial Condition

Statistical Prediction

$$\hat{T}_{m+\tau} = \hat{b}_{m,\tau} + \hat{a}_{m,\tau} T_m,$$

where $\hat{b}_{m,\tau}$ and $\hat{a}_{m,\tau}$ are least squares estimates of the slope and intercept estimated from independent data.

Monthly Mean NINO3.4 Forecasts by Regression
1982–1998 CLIM; lead= 2.5; alpha= 5%

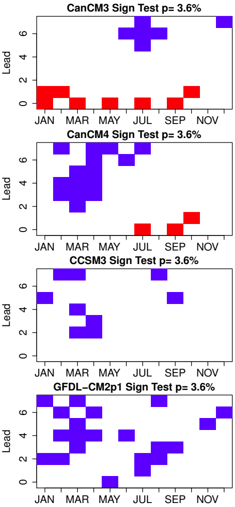


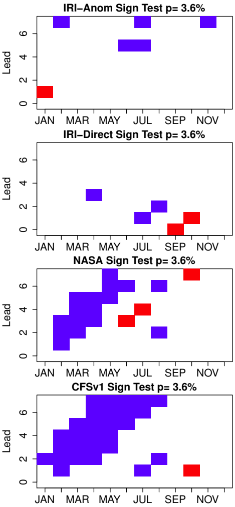
NMME Rankings for 1999-2014 NINO3.4 Hindcasts

1. CanCM3, Linear regression model
2. CanCM4
3. FLOR-A, FLOR-B, Multi-model mean
4. NASA, CM2p1-AER
5. CCSM3
6. CCSM4
7. CFSv2 (because of "dual climatology")

bias correction based on 1982-1998

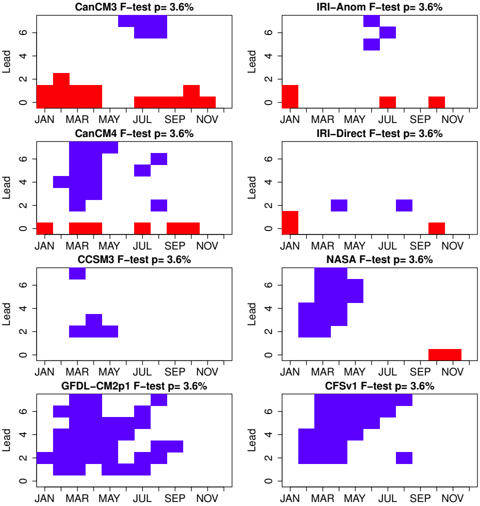
Compare to CFSv2 using Sign Test Based on MSE





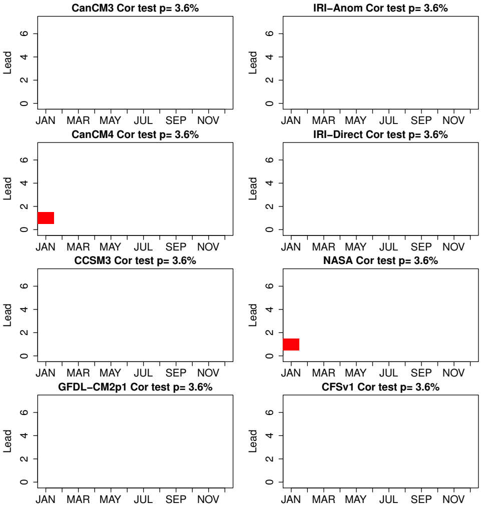
Blue = CFSv2 outperforms model. Red = Model outperforms CFSv2.

Compare to CFSv2 using F-test Based on MSE



Blue = CFSv2 outperforms model. Red = Model outperforms CFSv2.

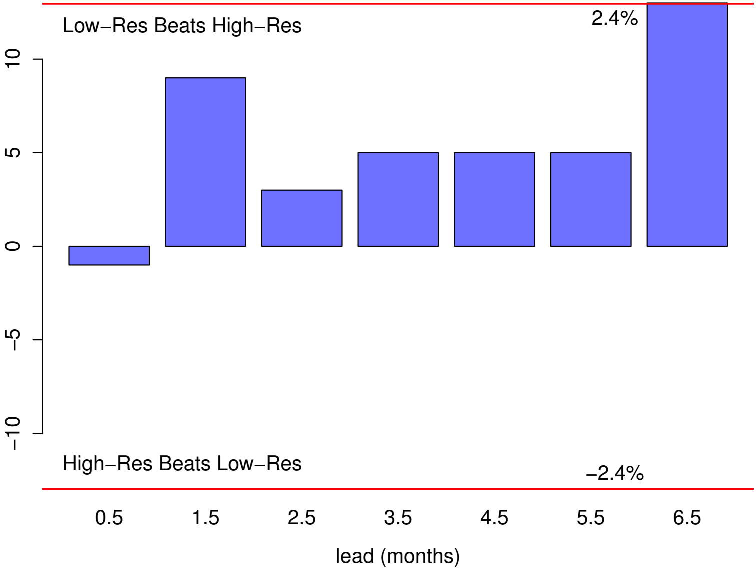
Compare to CFSv2 using Correlation Test



Blue = CFSv2 outperforms model. Red = Model outperforms CFSv2.

Minerva Compare T319 with T639

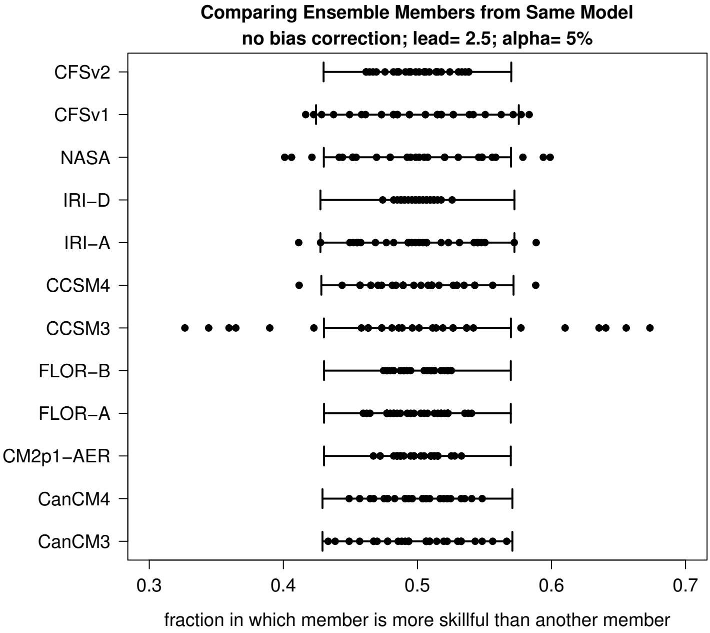
Only May and Nov ICs



Sum (sqerr(Hi−Res) > sqerr(Low−Res))
Nov IC, 1982−2010, 6 members, 1982−1994 CLIM

Exchangeability

Are the ensemble members distinguisable?
Compare skill of different ensemble members
from same model.


Comparing Ensemble Members from Same Model
no bias correction; lead= 2.5; alpha= 5%
CFSv2
CFSv1
NASA
IRI−D
IRI−A
CCSM4
CCSM3
FLOR−B
FLOR−A
CM2p1−AER
CanCM4
CanCM3
0.3
0.4
0.5
0.6
0.7
fraction in which member is more skillful than another member

Strictly Exchangeable

CanCM3: Different A-L-I-O initializations starting from different ICs

CanCM4: Different A-L-I-O initializations starting from different ICs

FLOR-A: Ensemble data assimilation

FLOR-B: Ensemble data assimilation

CM2p1-AER: Ensemble data assimilation

IRI-D: A-L initialized from AMIP runs

IRI-A: A-L initialized from AMIP runs

Not Strictly Exchangeable

NASA: some lagged ensemble, some breeding vectors

CCSM3: A-L-I initialized from different years in long control

CCSM4: Lagged ensemble for A, same I initialization as CCSM3

CFSv1: Lagged ensemble for A (more widely spaced than CFSv2)

CFSv2: Lagged ensemble for A-L

If some forecasts are better than others, then can we combine them to improve skill?

Model Diversity vs. Ensemble Size

$$\mathrm{SKILL}(V; F_1, F_2) > \mathrm{SKILL}(V; F_1)$$

Skill could be improved by

larger ensemble size	-or-	addition of new signals
reduction of noise		model diversity

Information Theory

$$\begin{array}{ccccc} \mathsf{SKILL}(V;F_1,F_2) & = & \mathsf{SKILL}(V;F_1) & + & \mathsf{SKILL}(V;F_2|F_1) \\ \text{Multimodel} & & \text{Single-model} & & \text{Conditional} \\ \text{Mutual Information} & & \text{Mutual Information} & & \text{Mutual Information} \end{array}$$

The condition for skill to increase by adding another forecast is

$$\text{Conditional Mutual Information} > 0$$

Gaussian Distributions

$$\mathrm{SKILL}(V; F_2|F_1) = -\frac{1}{2}\log\left(1 - \rho^2_{v2|1}\right)$$

$\rho_{v2|1}$ = partial correlation between V and F_2 conditioned on F_1:

If skill comes from reduction of noise, then

$$\rho^{\text{noise}}_{V2|1} \leq \sqrt{\frac{E_2}{(E_1 + E_2)(E_1 + 1)}}.$$

where E_1 and E_2 are the ensemble sizes of F_1 and F_2.

DelSole, Nattala, Tippett, 2014, Geophys. Res. Lett.

Statement of the Question

o: observations

f_1: forecast 1

f_2: forecast 2

- The skill of a forecast f_1 can be measured by correlation:

$$\rho = \operatorname{cor}[o, f_1]$$

- The skill of the best linear combination of f_1 and f_2 can be measured by the *multiple correlation*

$$R = \max_{\beta_1, \beta_2} \operatorname{cor}[o, \beta_2 f_1 + \beta_2 f_2]$$

Statement of the Question

o: observations

f_1: forecast 1

f_2: forecast 2

- The skill of a forecast f_1 can be measured by correlation:

$$\rho = \text{cor}\left[o, f_1\right]$$

- The skill of the best linear combination of f_1 and f_2 can be measured by the *multiple correlation*

$$R = \max_{\beta_1, \beta_2} \text{cor}\left[o, \beta_2 f_1 + \beta_2 f_2\right]$$

Question: Is $R > \rho$?

Statistical Test

$$\rho = \max_{\beta_1} \operatorname{cor}\left[o, \beta_1 f_1\right]$$

$$R = \max_{\beta_1,\beta_2} \operatorname{cor}\left[o, \beta_1 f_1 + \beta_2 f_2\right]$$

- The hypothesis $R = \rho$ is equivalent to the hypothesis $\beta_2 = 0$.
- Testing the hypothesis $\beta_2 = 0$ is standard and is based on

$$t = \frac{\hat{\beta}_2}{se_2}$$

- It can be shown that

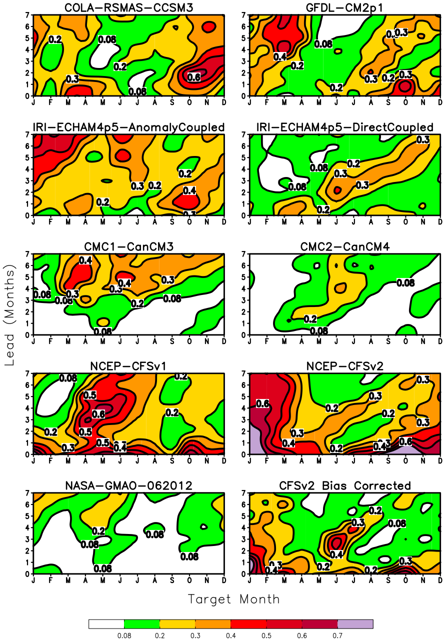
$$t^2 = \frac{R^2 - \rho^2}{1 - R^2} \frac{N-3}{1} = \frac{SSE_1 - SSE_{1+2}}{SSE_{1+2}} \frac{N-3}{1}$$

Equivalent Interpretation of the Test

A significant t-value means: if the forecast f_1 is regressed out of f_2, the residual forecast f_2' still has skill.

Combined Forecast

- We consider only equal weighting schemes.
- Equal weights is very competitive with more sophisticated schemes
    - Kharin and Zwiers, 2002, *J. Climate*
    - Hagedorn et al. 2005, *Tellus A*
    - Weigel et al. 2010, *J. Climate*
    - DelSole et al. 2012, *J. Climate*
    - Sansom et al. 2013, *J. Climate*



Improvement in NINO3.4 skill due to combining models

Conditional Mutual Information

5% significance ≥ 0.08

pure noise $\lesssim 0.22$

Summary

1. Skill measures computed on a common period or with a common set of observations are not independent.
2. Standard tests for differences in correlation or MSE are biased when evaluated over common period.
3. Random walk test avoids these problems and moreover applies to non-Gaussian distributions and arbitrary skill measures.
4. Canadian models are the most skillful dynamical models in NMME, even when compared to the multi-model mean.
5. A regression model is significantly more skillful than all but one dynamical model in the NMME (to which it is equally skillful).
6. There are significant skill differences between ensemble members from same model in NMME, reflecting differences from initialization.
7. Multimodel ensembles have higher skill than any single model, and this increase is due to model diversity, not increased ensemble size.